

# GPU Programming

## A taxonomy of hardware parallelism

Christian Lessig

# Parallel programming

“[Serial] algorithms have improved faster than clock over the last 15 years. [Parallel] computers are unlikely to be able to take advantage of these advances because they require new programs and new algorithms.”

Gordon Bell (1992)

G. Bell, “Massively parallel computers: why not parallel computers for the masses?,” in *The Fourth Symposium on the Frontiers of Massively Parallel Computation*, 1992, pp. 292–297.

# Parallel programming

“[Serial] algorithms have improved faster than clock over the last 15 years. [Parallel] computers are unlikely to be able to take advantage of these advances because they require new programs and new algorithms.”

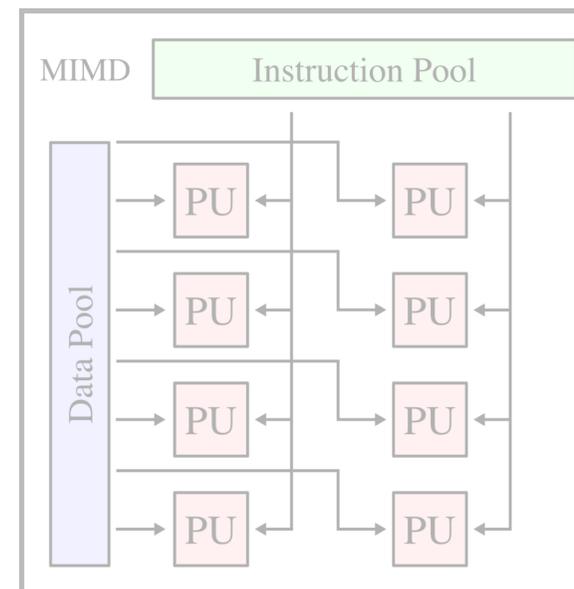
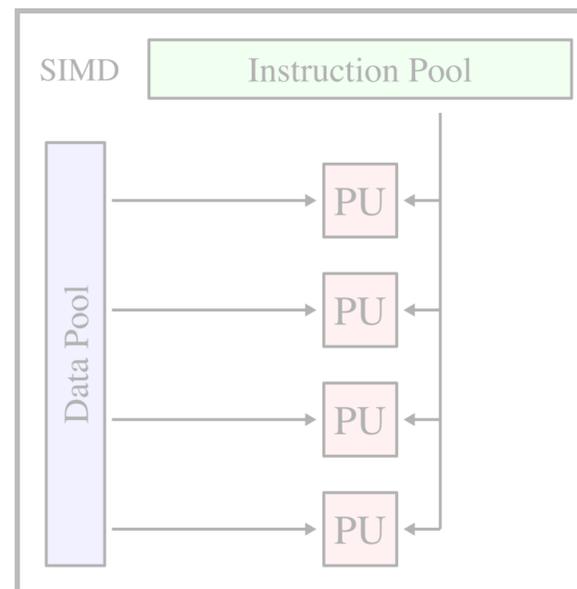
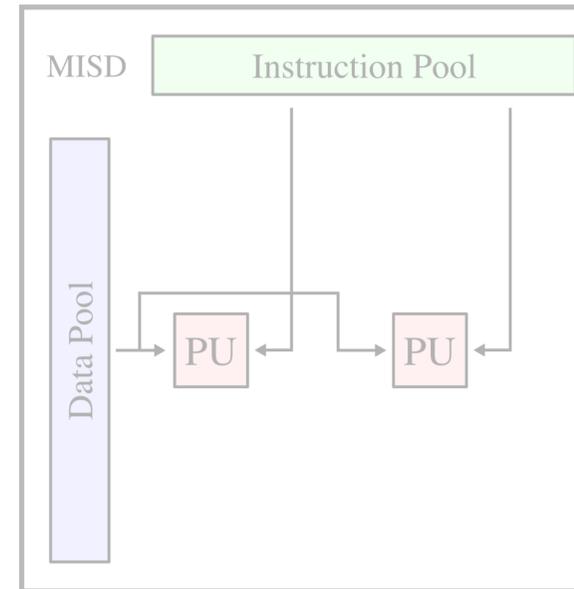
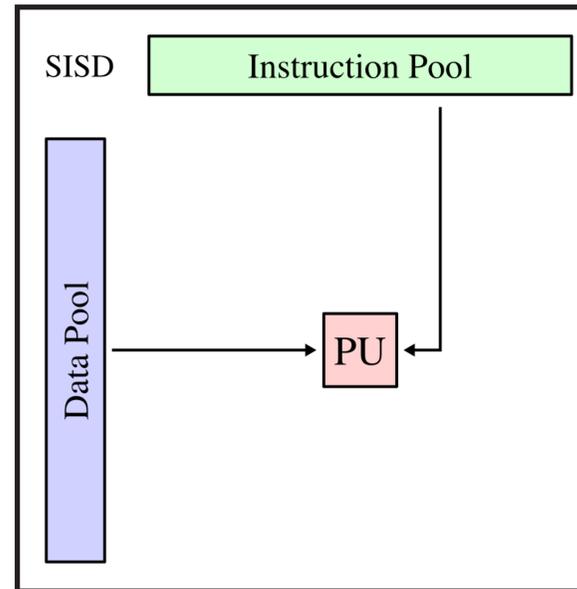
**Even worse:  
You have to know your  
architecture!**

Gordon Bell (1992)

G. Bell, “Massively parallel computers: why not parallel computers for the masses?,” in *The Fourth Symposium on the Frontiers of Massively Parallel Computation*, 1992, pp. 292–297.

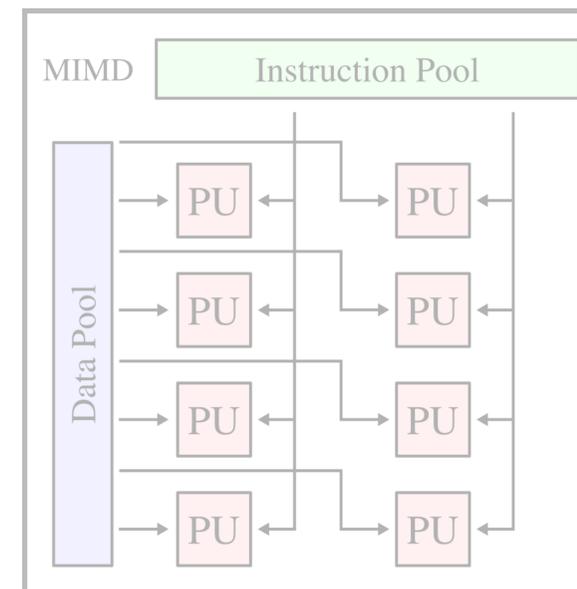
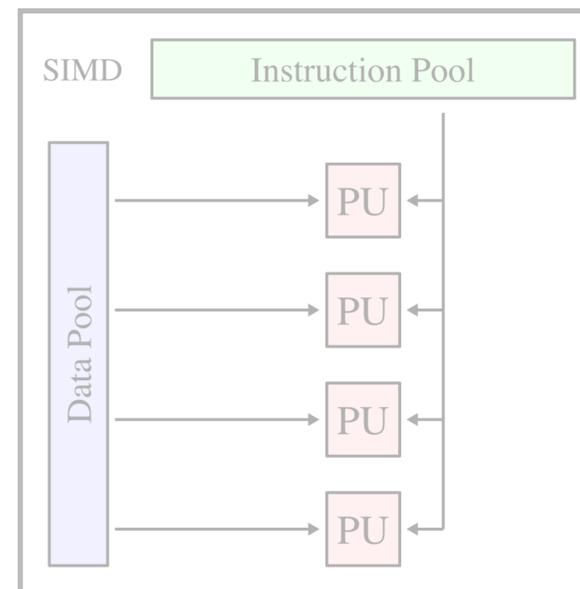
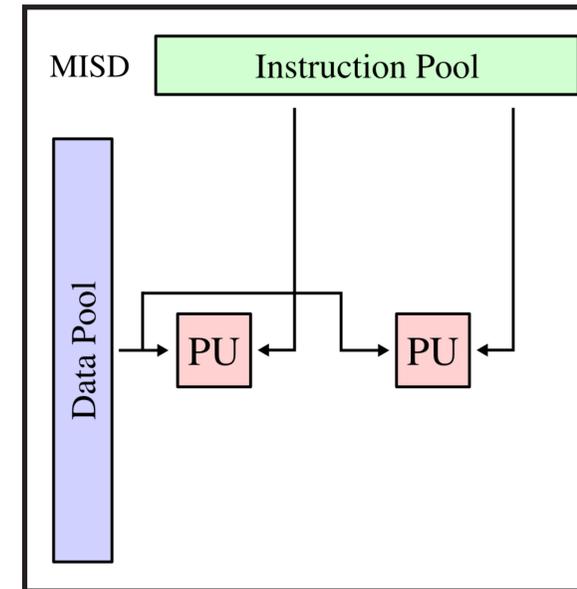
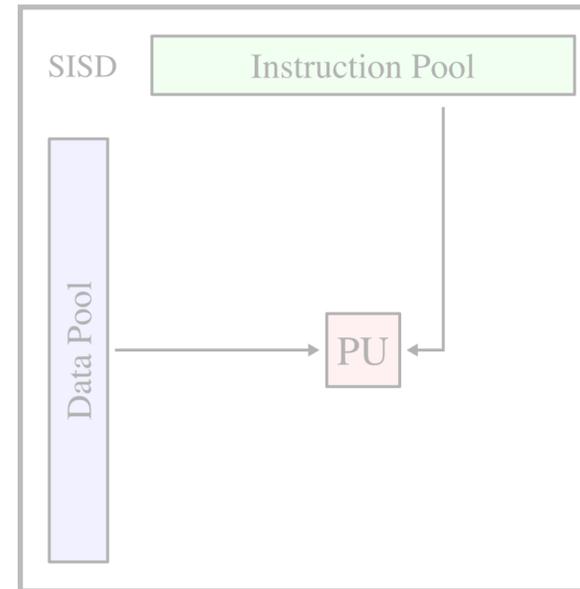
# Flynn's classification

# Flynn's classification



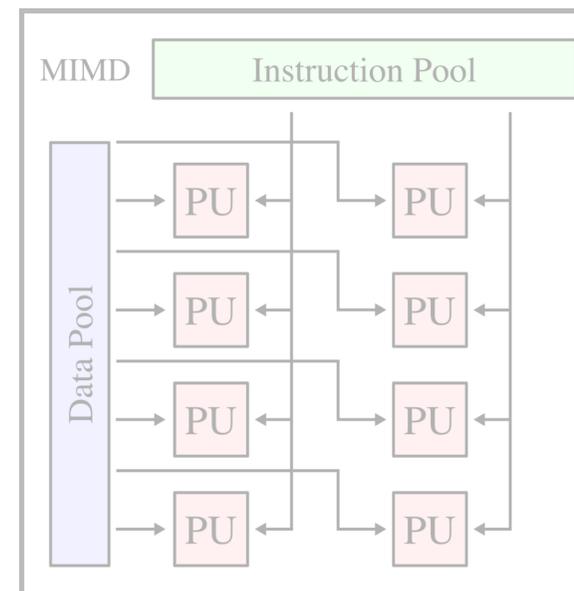
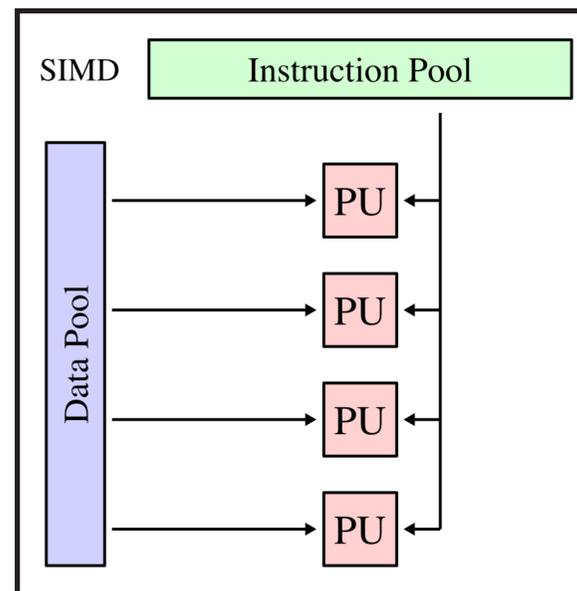
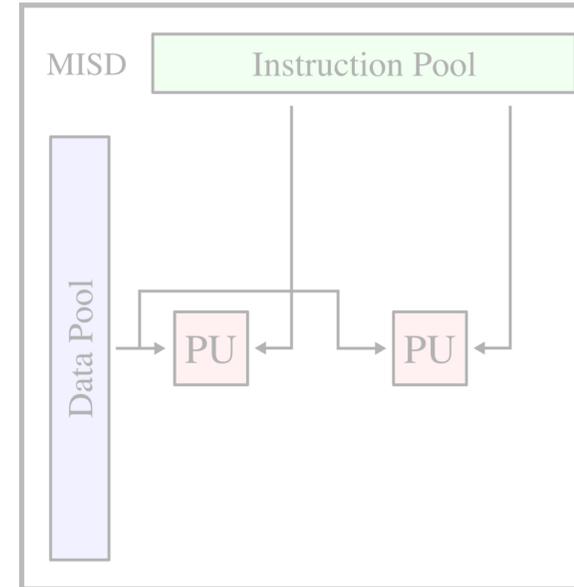
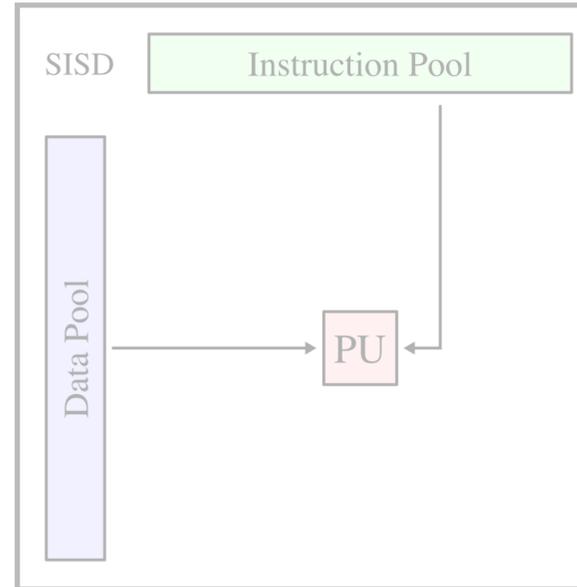
[https://en.wikipedia.org/wiki/Flynn%27s\\_taxonomy](https://en.wikipedia.org/wiki/Flynn%27s_taxonomy)

# Flynn's classification



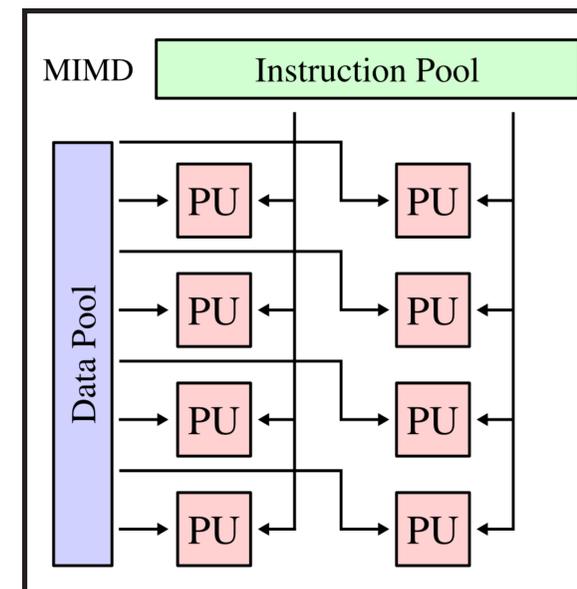
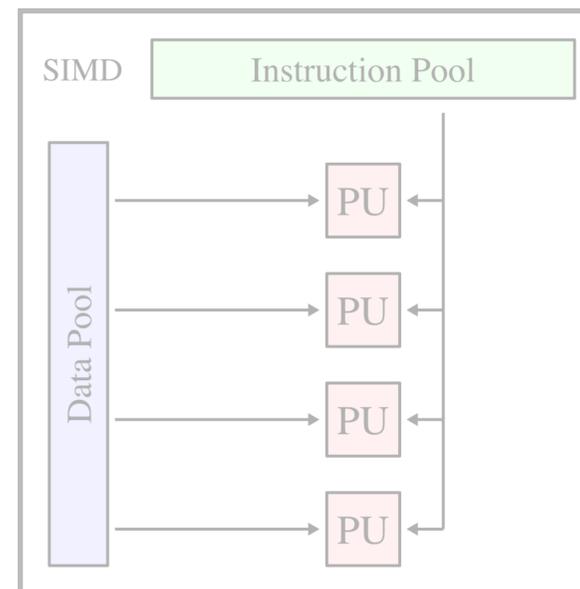
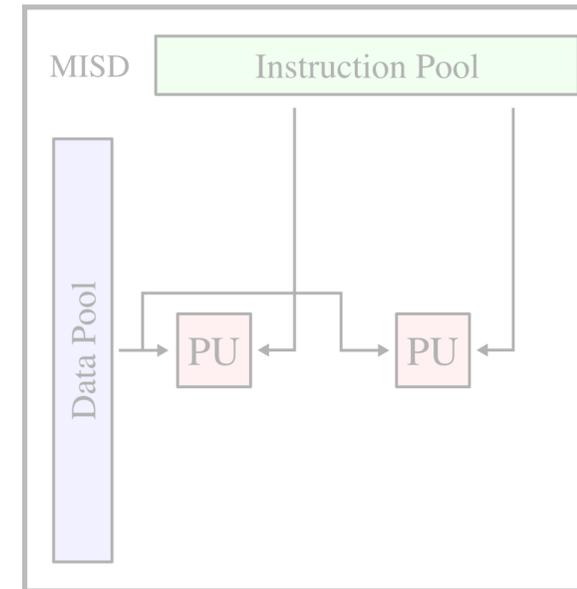
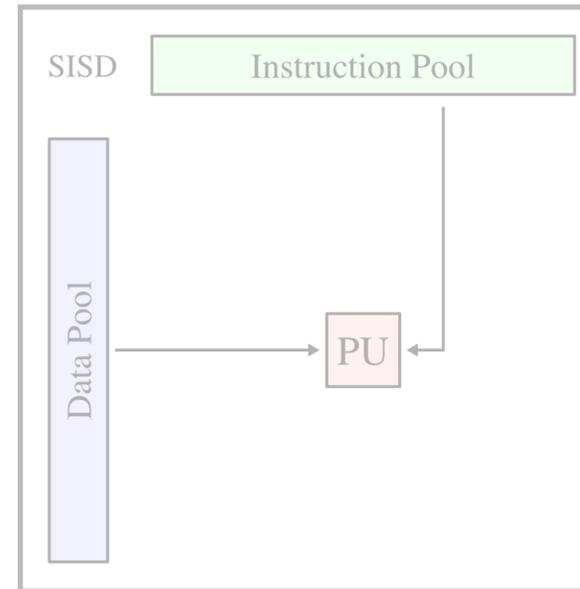
[https://en.wikipedia.org/wiki/Flynn%27s\\_taxonomy](https://en.wikipedia.org/wiki/Flynn%27s_taxonomy)

# Flynn's classification



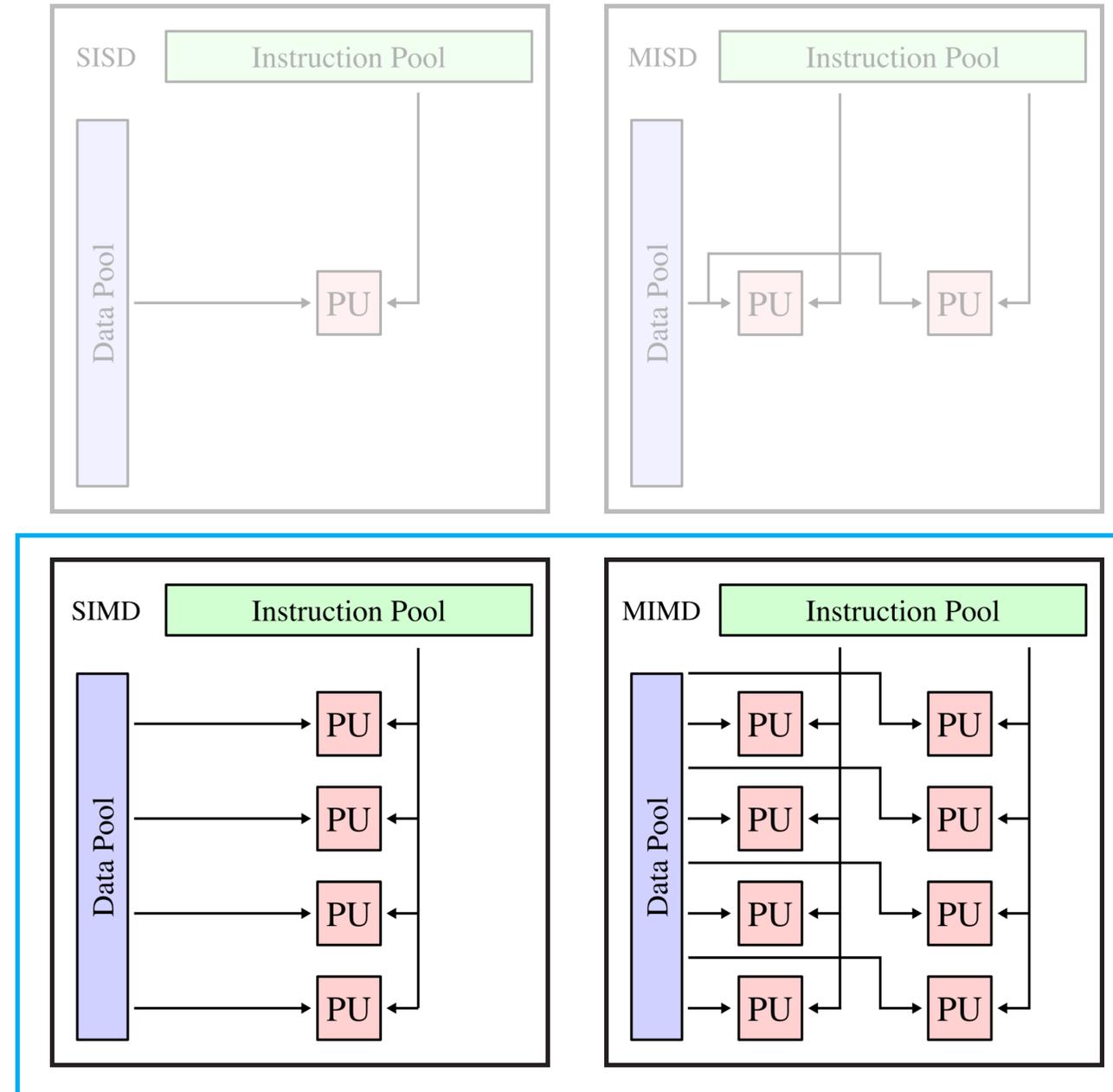
[https://en.wikipedia.org/wiki/Flynn%27s\\_taxonomy](https://en.wikipedia.org/wiki/Flynn%27s_taxonomy)

# Flynn's classification



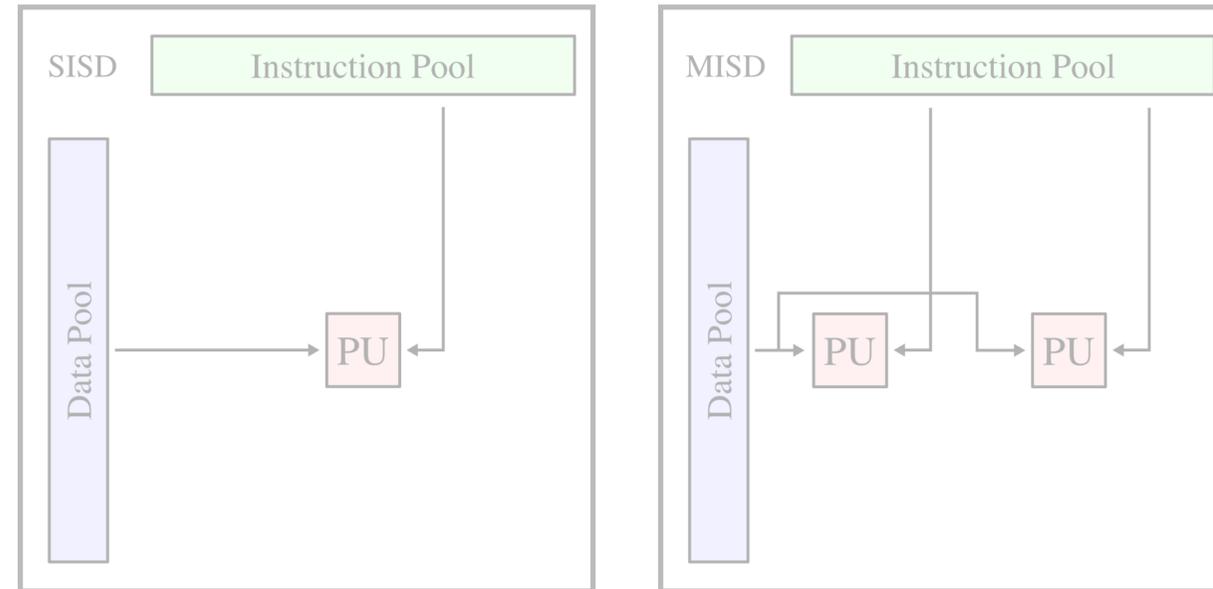
[https://en.wikipedia.org/wiki/Flynn%27s\\_taxonomy](https://en.wikipedia.org/wiki/Flynn%27s_taxonomy)

# Flynn's classification

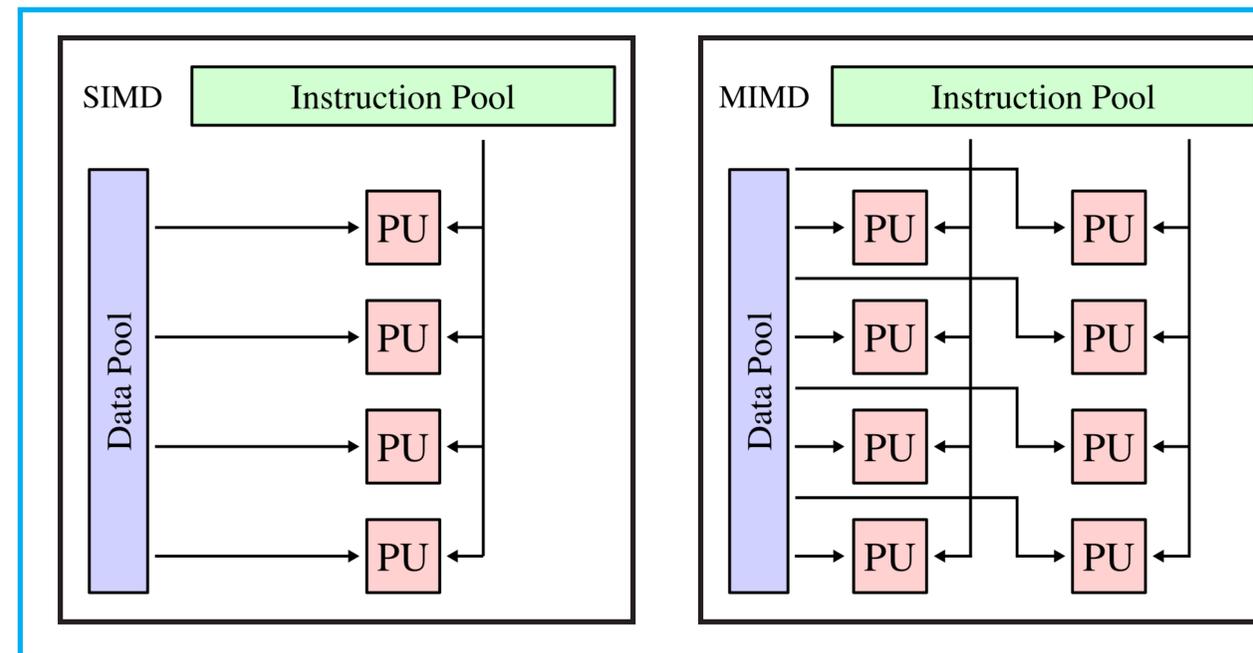


existing  
parallel  
architectures

# Flynn's classification



data  
parallelism

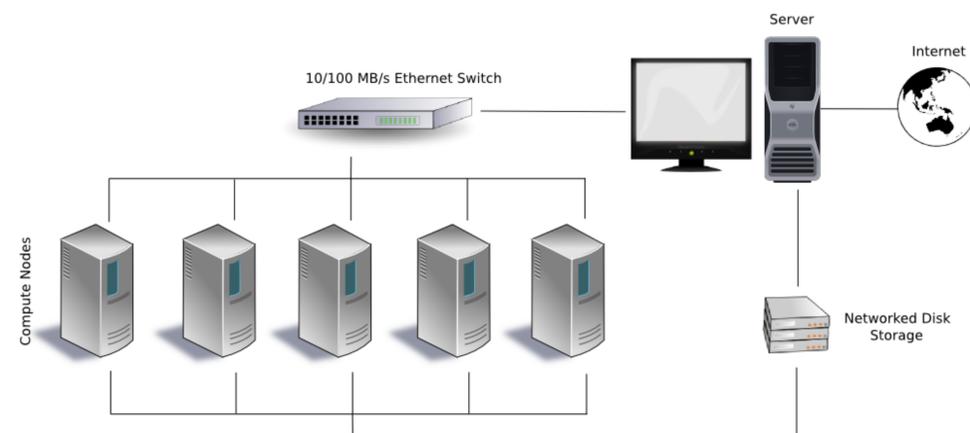


task  
parallelism

# Classification according to memory

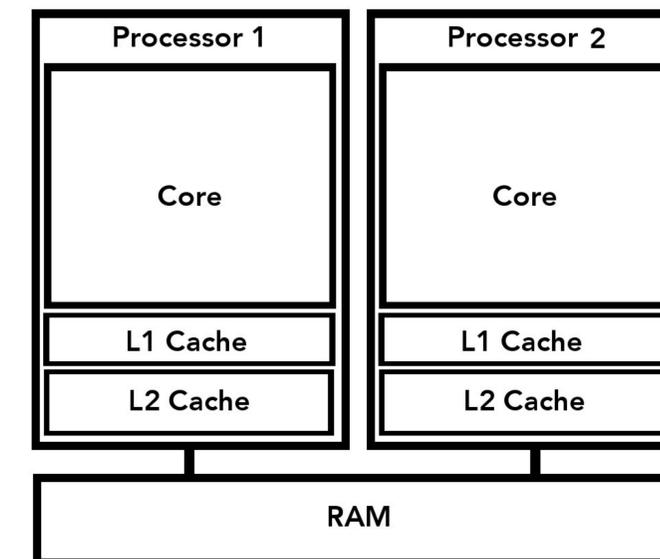
# Classification according to memory

## Distributed memory



<https://upload.wikimedia.org/wikipedia/commons/4/40/Beowulf.png>

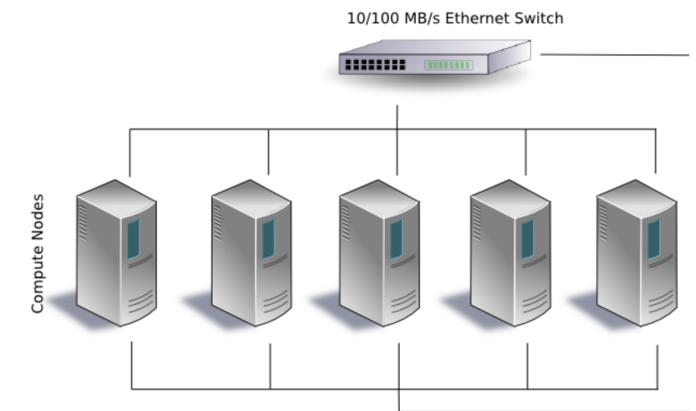
## Shared memory



[https://en.wikipedia.org/wiki/Multi-core\\_processor](https://en.wikipedia.org/wiki/Multi-core_processor)

# Distributed memory

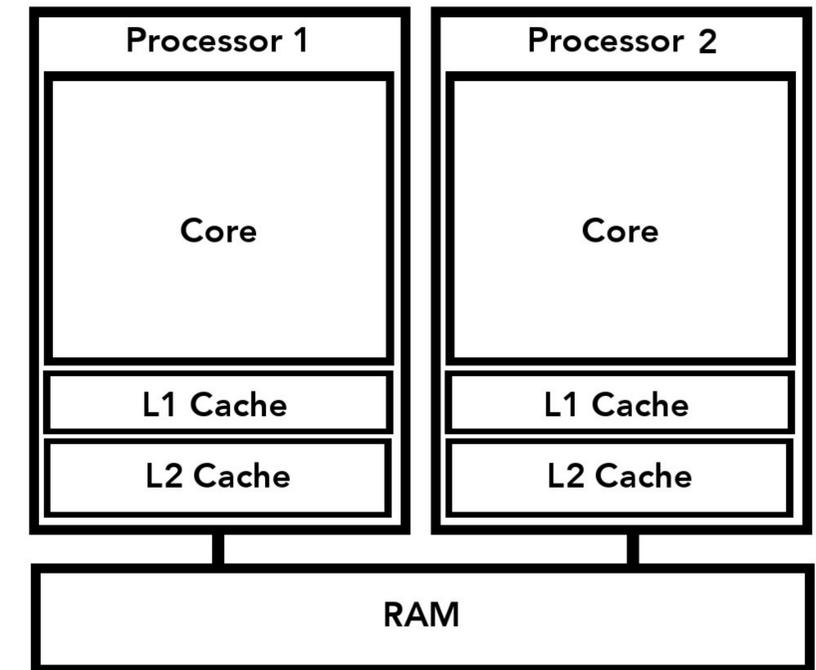
- Scales to arbitrary number of processors
- Communication via message passing
  - › e.g. MPI
  - › High latency and limited bandwidth
- Used in super-computers



<https://upload.wikimedia.org/wikipedia/commons/4/40/Beowulf.png>

# Shared memory

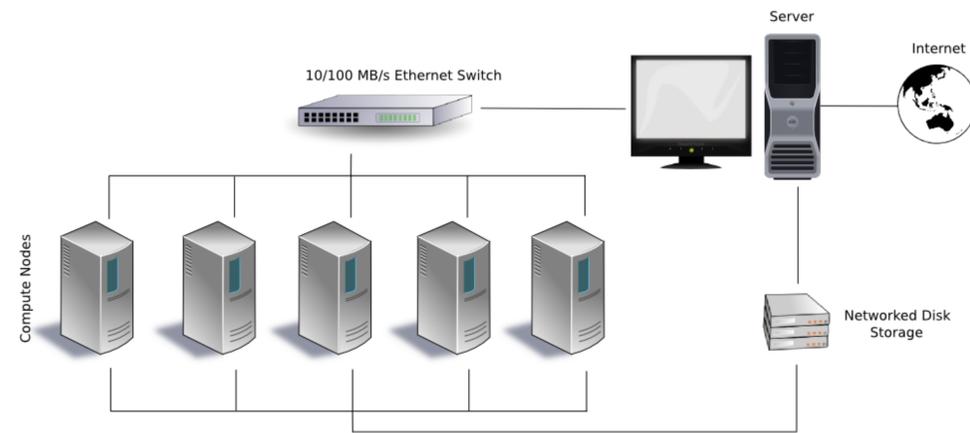
- Limited to less than  $\approx 100$  processors
- Communication via “shared” memory
  - › Low latency and high bandwidth
  - › Access to shared resources needs to be synchronized



[https://en.wikipedia.org/wiki/Multi-core\\_processor](https://en.wikipedia.org/wiki/Multi-core_processor)

# Classification according to memory

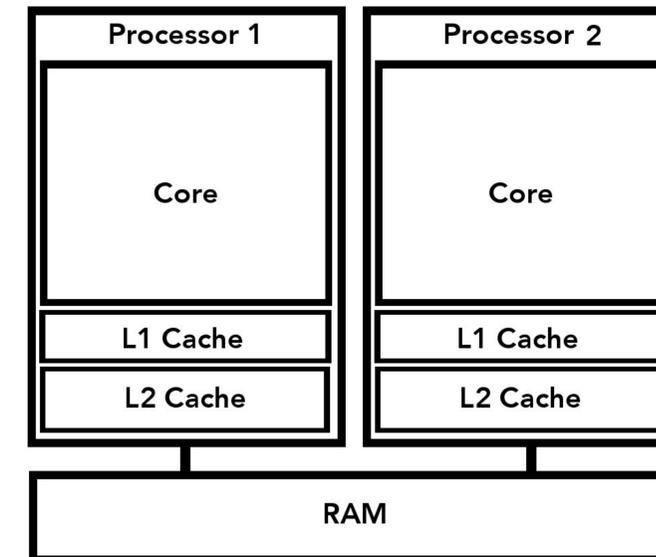
## Distributed memory



<https://upload.wikimedia.org/wikipedia/commons/4/40/Beowulf.png>

## Message passing

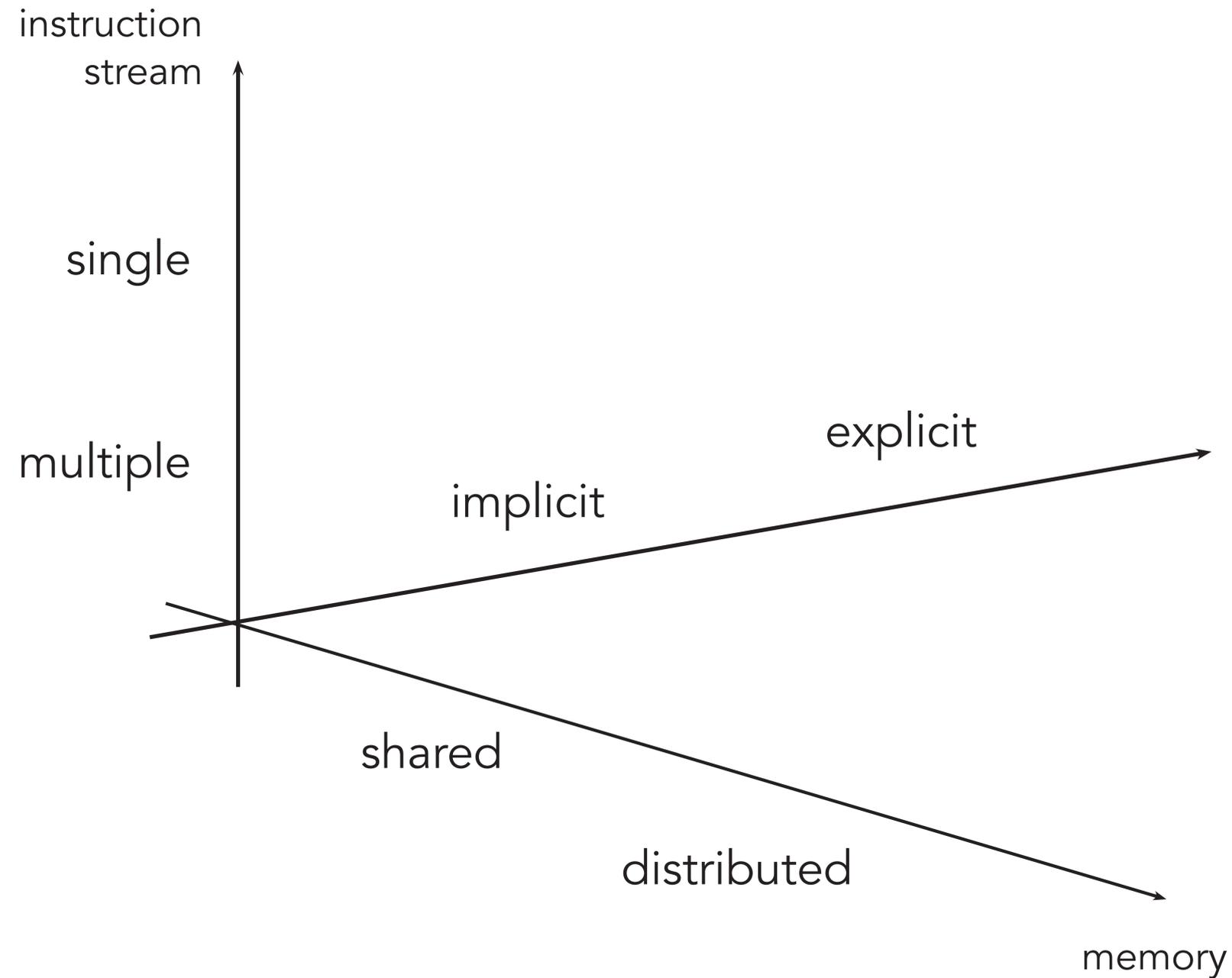
## Shared memory



[https://en.wikipedia.org/wiki/Multi-core\\_processor](https://en.wikipedia.org/wiki/Multi-core_processor)

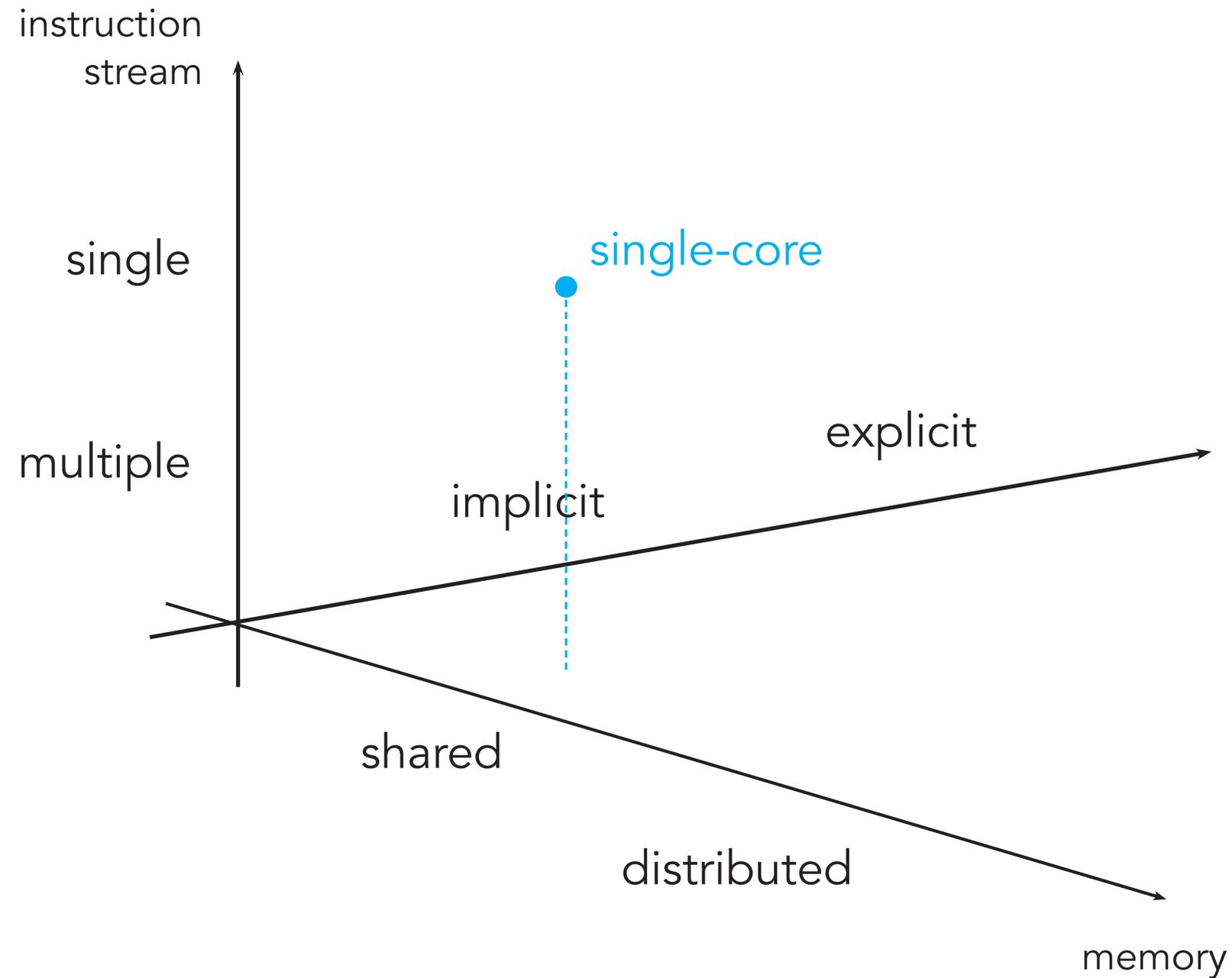
## Multi-threading

# Taxonomy



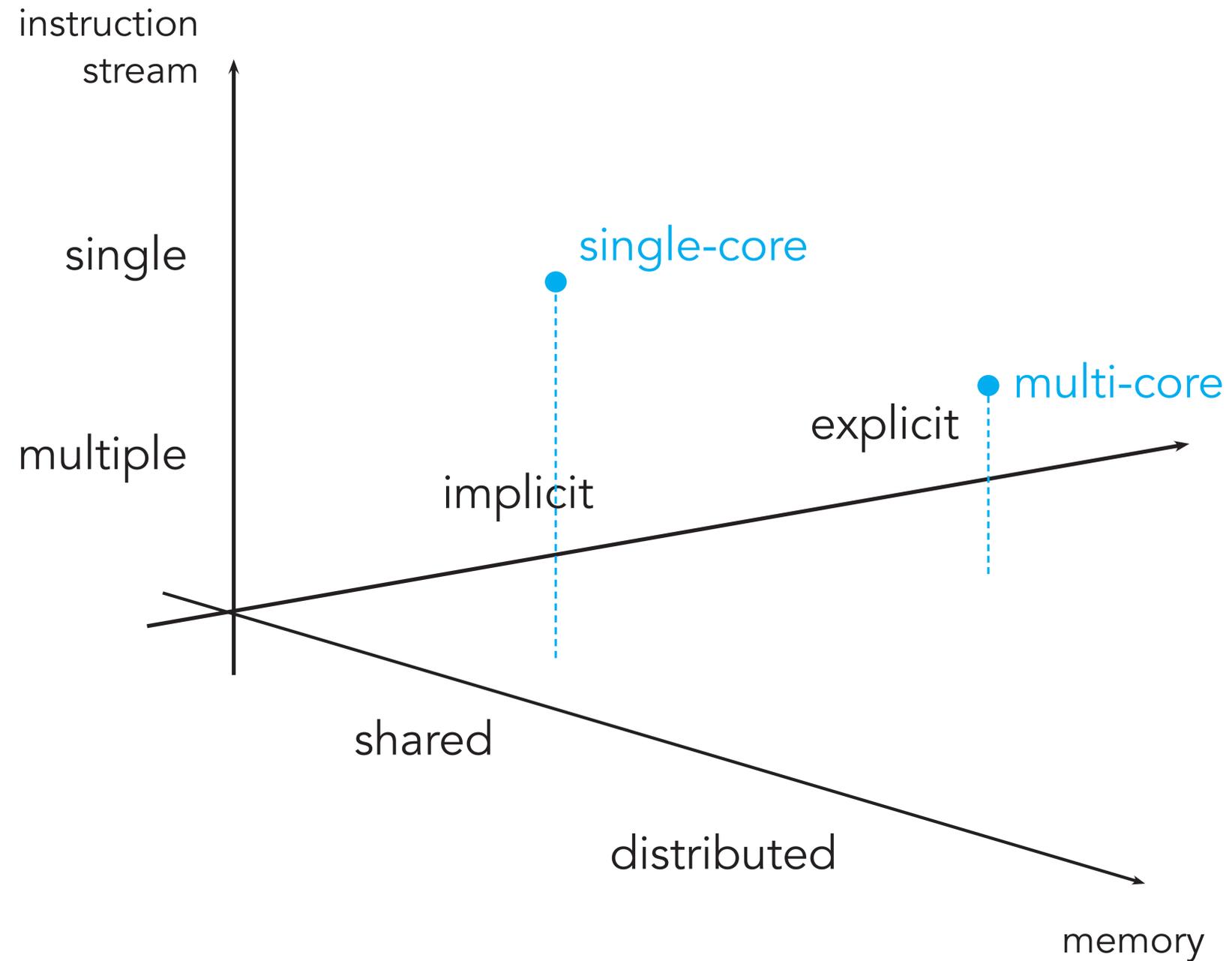
- cluster
- multi-core
- single-core
- gpu

# Taxonomy



- cluster
- multi-core
- gpu

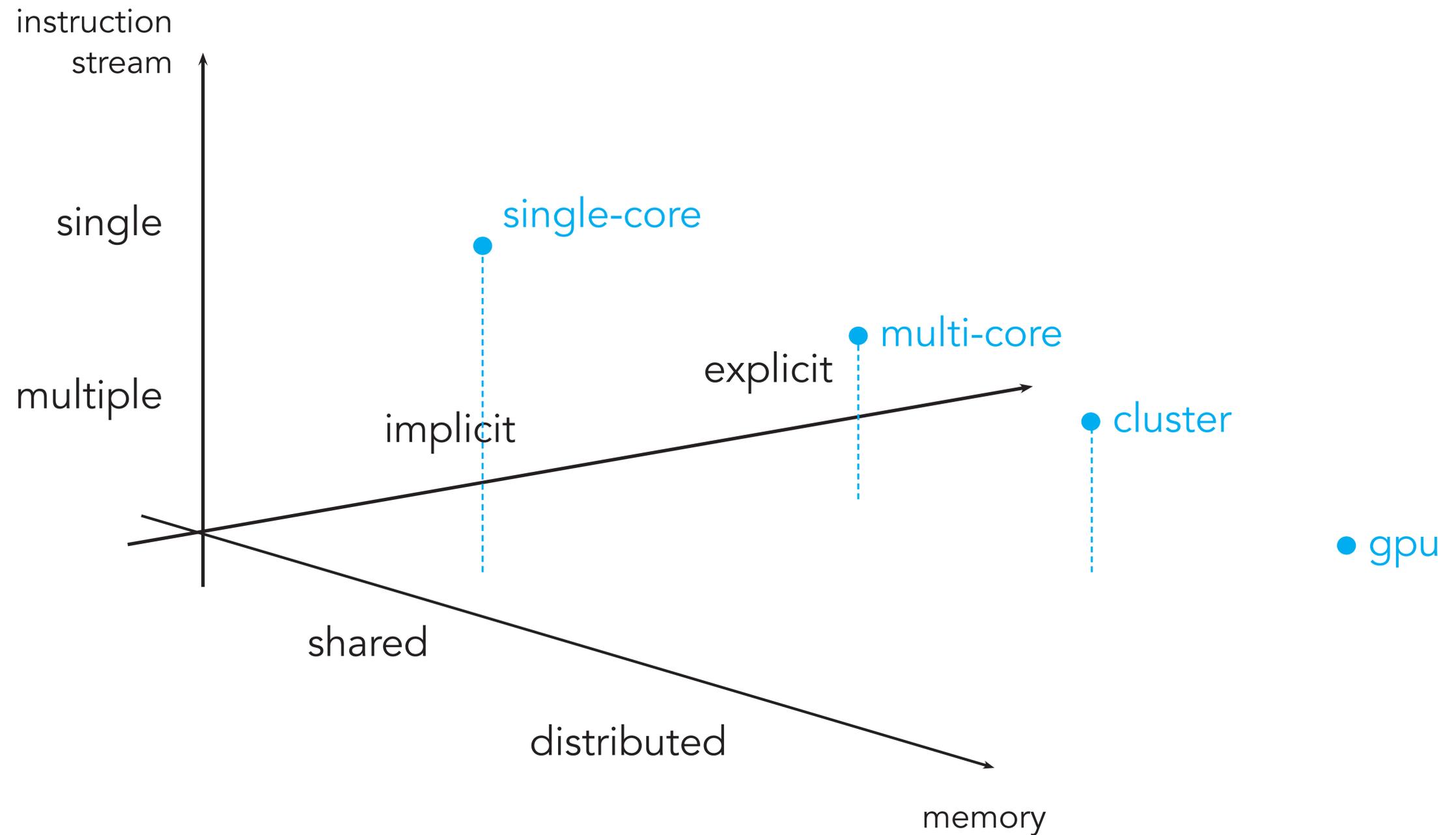
# Taxonomy



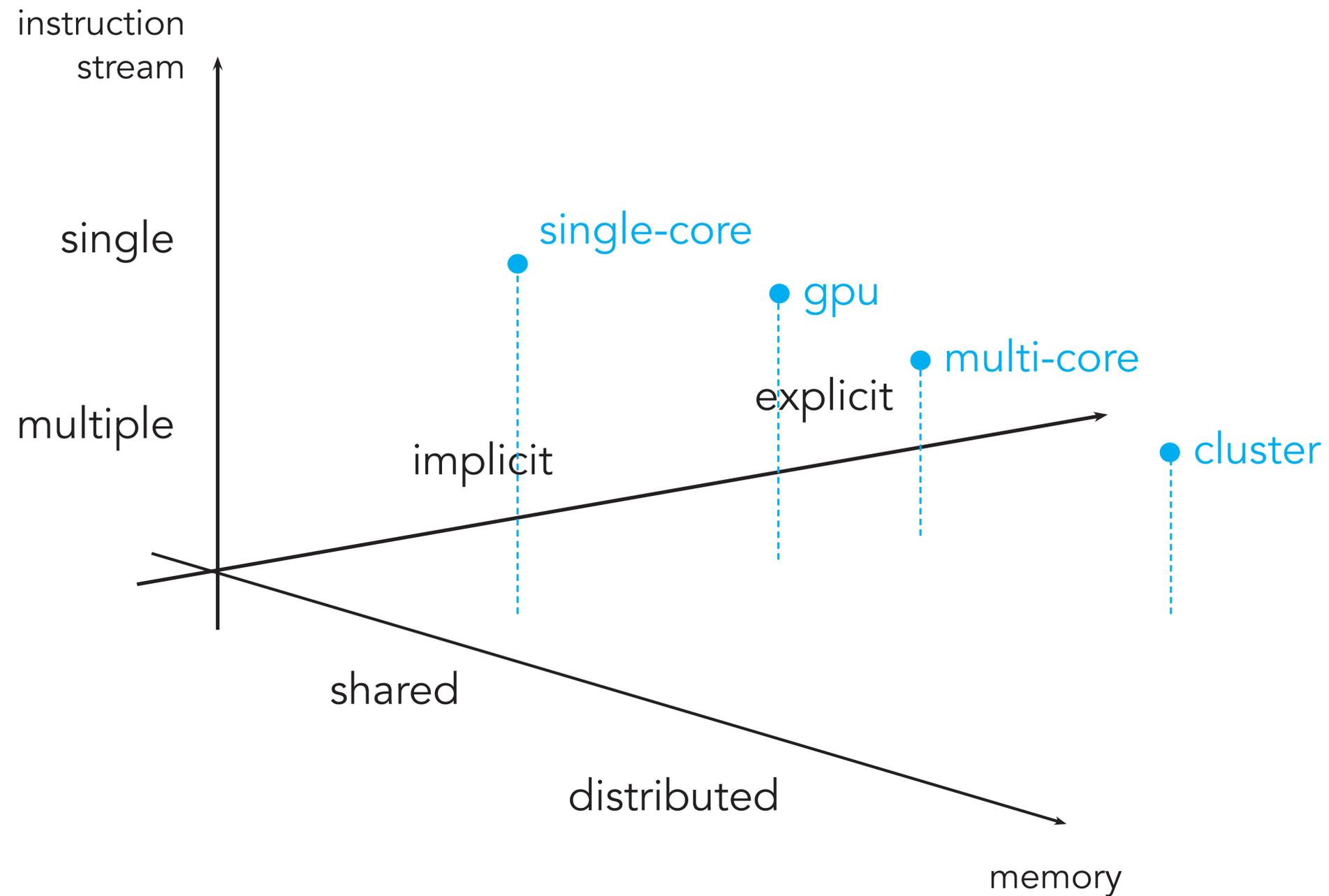
● cluster

● gpu

# Taxonomy



# Taxonomy



# Further reading

- J. L. Hennessy and D. A. Patterson, Computer architecture: a quantitative approach, fourth edition. Morgan Kaufmann, 2007.
- <http://cva.stanford.edu/classes/cs99s/>
- <http://research.ac.upc.edu/HPCseminar/SEM9900/Pollack1.pdf>
- <http://groups.csail.mit.edu/cag/raw/documents/Waingold-Computer-1997.pdf>
- <http://cacm.acm.org/magazines/2009/5/24648-spending-moores-dividend/fulltext>